Quality Assessment of Domesticated Animal Genome Assemblies

Seemann, Stefan E; Anthon, Christian; Palasca, Oana; Gorodkin, Jan

Published in:
Bioinformatics and Biology Insights

DOI:
10.4137/BBI.S29333

Publication date:
2015

Document version
Publisher's PDF, also known as Version of record

Quality Assessment of Domesticated Animal Genome Assemblies

Stefan E. Seemann, Christian Anthon, Oana Palasca and Jan Gorodkin

Center for non-coding RNA in Technology and Health, Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, København, Denmark.

Supplementary Issue: Current Developments in Domestic Animal Bioinformatics

ABSTRACT: The era of high-throughput sequencing has made it relatively simple to sequence genomes and transcriptomes of individuals from many species. In order to analyze the resulting sequencing data, high-quality reference genome assemblies are required. However, this is still a major challenge, and many domesticated animal genomes still need to be sequenced deeper in order to produce high-quality assemblies. In the meanwhile, ironically, the extent to which RNAseq and other next-generation data is produced frequently far exceeds that of the genomic sequence. Furthermore, basic comparative analysis is often affected by the lack of genomic sequence. Herein, we quantify the quality of the genome assemblies of 20 domesticated animals and related species by assessing a range of measurable parameters, and we show that there is a positive correlation between the fraction of mappable reads from RNAseq data and genome assembly quality. We rank the genomes by their assembly quality and discuss the implications for genotype analyses.

KEYWORDS: genome assembly, domesticated animals, assembly quality

SUPPLEMENT: Current Developments in Domestic Animal Bioinformatics

CITATION: Seemann et al. Quality Assessment of Domesticated Animal Genome Assemblies. *Bioinformatics and Biology Insights* 2015:9(S4) 49–58 doi: 10.4137/BBI.S29333.

TYPE: Original Research

RECEIVED: October 19, 2015. **RESUBMITTED:** May 02, 2016. **ACCEPTED FOR PUBLICATION:** May 03, 2016.

ACADEMIC EDITOR: J. T. Efird, Associate Editor

PEER REVIEW: Nine peer reviewers contributed to the peer review report. Reviewers' reports totaled 3,209 words, excluding any confidential comments to the academic editor.

FUNDING: This study was funded by the Danish Center for Scientific Computing (DCSC, DeiC); Innovation Fund Denmark (Programme Commission on Strategic Growth Technologies); and The Danish Council for Independent Research. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: seemann@rth.dk

Introduction

Domesticated farm animals are of the highest importance for human food supply. This implies a need for optimized productivity while demanding healthy animals living under justifiable ethical conditions. Some of the domestic animals are also used as model organisms for human diseases, eg, pig as a model for obesity, cardiovascular disease, gastroenteropathy, and immunological diseases, as well as a pharmacology and toxicology model.[1–6] Variations in the genomic sequence are gaining increasing importance for improving strategies for domestic animal studies. However, the assembled genomes are of highly diverse assembly quality. High-quality genome assemblies are a prerequisite for high-quality genomic and transcriptomic analyses, while in contrast, poor genome assembly qualities increase the risk of poor transcriptome assemblies, which highly impact the value of any next-generation sequencing (NGS) experiment. In recent years, various NGS strategies are widely used to address a wide range of different questions from differential expression to epigenetic marks such as methylation signatures of RNA transcripts. Although the creativity in the ways to use NGS seems to be unlimited, the usage of NGS often requires a good reference genome to start with. Unfortunately, there are seemingly not many recent advances in improving the reference genome sequences accordingly, although ongoing development in the field such as PacBio holds the potential for a paradigm shift, pending on the overall costs.[7] Currently, genomes such as pig (susScr10.2) and dog (canFam3) have not been improved since 2011, and the horse genome has not been improved since 2007. This leaves an apparent imbalance and a potential waste of resources by generating the data meant for genome-wide comparison that cannot be mapped. It also influences proper and full analyses. In the best case, this will result in an incomplete analysis, but in the worst case, it will lead to misinterpretation of the data.

We briefly outline some of the genome assemblies in Figure 1A. Most of the species considered were sequenced using a hybrid approach, combining whole-genome shotgun sequencing (WGS) with a hierarchical BAC clone approach, and only a few of them solely relied on WGS (for example, dog, sheep, and goat). Organisms sequenced in the early 2000s benefit from the integration of Sanger-based sequencing, which is characterized by longer read length and better sequence quality, while more recently, sequenced organisms are mainly based on short-read NGS.[8] The dog genome assembly is based on WGS with Sanger sequencing. CanFam2, on which the current canFam3 is based, was at its

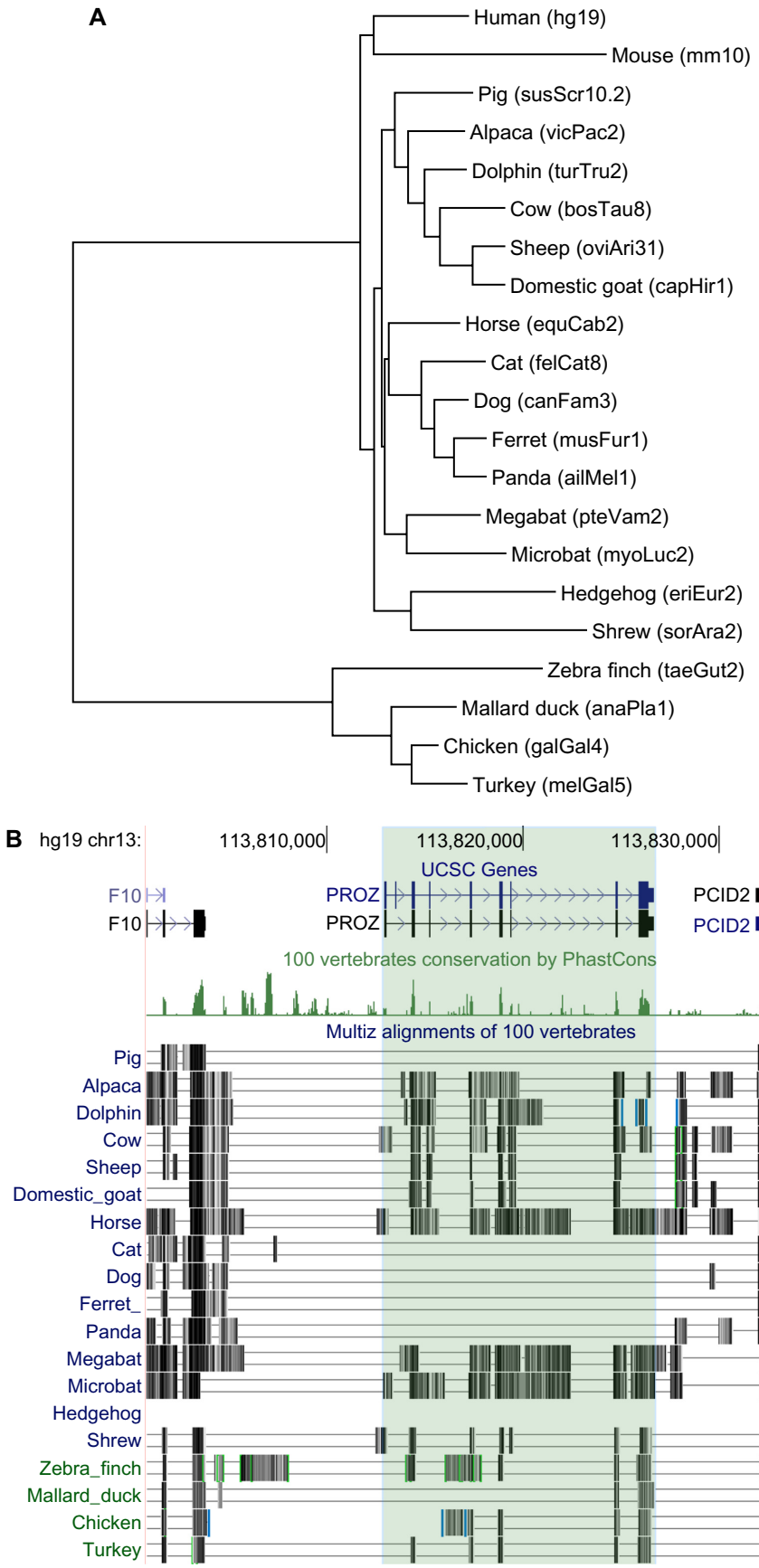

Figure 1. Discrepancy between phylogeny and gene annotation.

Notes: (**A**) The phylogenetic tree of the 21 investigated species is shown, with a clear separation between placental mammals and birds. The tree is a subset of the UCSC-generated 100-way tree. (**B**) A UCSC genome browser view in human of the genomic region around PROZ. PROZ is missing in the pig assembly susScr102 and in the phylogenetic subtree around dog, but the gene is conserved in the phylogenetic subtree of pig and even in the more distant birds.

time of release (2004) much better than most other assemblies (eg, the mouse genome), due to high sequence coverage and good data quality in terms of read length or library insert sizes (up to 200 kb insert length).[9] The horse genome, also known to be of high quality, is based on Sanger sequencing as well, but supplemented with BAC and fosmid clone maps, for better contiguity.[10] The cow genome was sequenced using the Sanger method and incorporates BAC clones, accounting for a large proportion of the genome coverage.[11] In cow, special attention was given to the genome assembly method, by using improved postprocessing algorithms that took into account synteny with the human genome, among others.[11] On the other hand, the pig genome is mainly based on sequences from BAC clones obtained back in the 2000s, to which WGS Illumina sequences are added for resolving gaps.[12]

The sheep and goat assemblies were obtained by short-read NGS. The sheep genome was iteratively improved with new sequencing data produced in different rounds, produced by Illumina and 454 technologies, in an attempt to cover the numerous gaps in the assembly.[13] For the goat assembly based on Illumina reads, the newer optical mapping technology was employed,[14] instead of radiation-hybrid or fluorescence in situ hybridization maps used in most other assemblies for the alignment of chromosomes. The chicken genome was initially sequenced in 2004 using the Sanger technology[15] with the support of BAC clone physical maps for assembly scaffolding. The first version was updated with new 454 sequencing data, genetic maps, and better assembly algorithms. The genome assembly of turkey was obtained using Illumina and 454 sequencing and makes use of a BAC clone-based physical map for aligning to chromosomes only.[16] Being less than half the size of mammalian genomes, but with a higher number of chromosomes, both chicken and turkey genomes showed particular assembly challenges in certain genomic regions, mainly due to repeats specific to the small chromosomes in birds.[15,16]

When conducting comparative genome analysis of specific regions of domesticated animal genomes in, eg, the University of California Santa Cruz (UCSC) genome browser, it is apparent that regions with some frequency are missing. For example, vitamin K-dependent plasma glycoprotein (PROZ), a gene encoding for a protein with a role in regulating blood coagulation in human, has annotated orthologs in cow, sheep, and horse, as well as mouse, and even xenopus and some birds, but orthologs are missing in pig and dog, among others (see Fig. 1B). However, it is not clear if this gene is missing because it is simply not in the genome assemblies or whether it was indeed lost in some of the lineages. Elucidation of these loci is highly relevant, so that it can be determined why the gene is missing in the genome. For example, if the missing genome sequence is a protein-coding gene highly conserved over mammals, one would also expect to find it in the syntenic region and a naive first approach is to look for the corresponding protein isoforms in the relevant databases. However, less conserved genomic loci will leave further analyses open for future considerations, or they would demand large resources to solve a local genomic region experimentally.

To investigate the extent of this problem, we assessed the quality of the most recent genome assemblies of 20 domestic animals and related species and compared to the human genome (hg19), which we define as the gold standard assembly. The assembly quality is measured considering a range of parameters, as follows: nucleic acid conservation of highly conserved protein-coding and ultraconserved elements (UCs); amino acid homology of universal single-copy orthologs; structure conservation of housekeeping RNAs; assembly sequence quality; and assembly contiguity. With this information, we can further quantify the imbalance between NGS applications and genome assembly quality.

Methods

Genomes. The genomes investigated in this study are listed in Table 1. We focused on domesticated animals that are part of either the Laurasiatheria or the Aves. We supplemented the domesticated animal genomes with other species within these two phylogenetic classes based on the criteria of maturity of the assembly and the existence of genomic annotation. For comparison, the well-assembled genomes of human (hg19/GRCh37) and mouse (mm10/GRCm38) were added. The genomic sequence was downloaded from the UCSC webservers as so-called 2 bit files.

Genome assembly quality features. Genome assembly quality has previously been assessed in many different ways with focus on methodologies (eg, insert size distributions and sequence coverage), genome biases (eg, *k*-mer distributions), or fragment length distributions (eg, N50).[17,18] Additionally, the completeness of highly conserved orthologous genes in genome assemblies has been investigated to reflect the expected gene content.[19,20] In the current study, we combine a number of these previously proposed features along with nucleic acid conservation and synteny of highly conserved genomic loci.

Analysis of conserved genomic features. The analysis of highly conserved genomic features (conserved protein-coding genes and UCs) is based on pairwise sequence alignments of human and the 20 vertebrates. The pairwise alignments were built by `lastz`[21] and the UCSC toolkit[22] for chains and nets with human as query. We used the UCSC tool `liftOver` (parameter minMatch = 0.8) to convert genomic coordinates in human to the other species based on the pairwise alignments. We investigated the conservation of the union of 32 universal genes (COGs) described by Ciccarelli et al.[23] and 444 conserved core eukaryotic genes[19] with an ortholog in human. The majority of COGs are ribosomal proteins. The 4,856 merged exons from these 463 protein-coding genes were classified as deleted, partially deleted, split, or being in the wrong order compared with the exon order in human. Furthermore, we checked for the presence of 473 UCs (200 bp long loci of 100% identity in rat, mouse, and human),[24] which were classified as deleted, partially deleted, or split. Note that

Table 1. Genome assembly quality features of human, domesticated animals, and related species.

SPECIES	ASSEMBLY			PROTEIN-CODING (PCE)				ULTRACONSERVED (UC)			ORTHOLOGS (BUSCO)				rRNA	tRNA	GAPS	CONTIGUITY (N50)
				4,856 EXONS				473 LOCI			3,023 GENES					21 AA		
	VERSION	YEAR	SIZE [MBP]	D [#]	P [#]	S [#]	W [#]	D [#]	P [#]	S [#]	C [%]	CD [%]	F [%]	M [%]	[SCORE]	[#]	[#]	[KBP]
Human	hg19	2009	3,137	0	0	0	0	0	0	0	90	1.7	5.1	4.5	8	20	411	46,396
Laurasiatheria																		
Mouse	mm10	2012	2,731	17	23	2	5	0	0	0	91	2.2	4.8	3.8	3	21	582	52,589
Panda	ailMel1†	2009	2,300	30	34	11	44	0	0	0	88	0.5	8.0	3.4	0	21	108,147	1,282
Cow	bosTau8	2009	2,670	23	26	6	24	3	1	0	84	1.8	8.6	6.9	6	21	72,051	6,380
Dog	canFam3	2011	2,411	33	22	9	24	6	2	0	89	2.0	6.3	4.2	8	20	23,876	45,877
Domestic goat	capHir1	2013	2,636	126	91	23	79	3	1	0	79	1.0	12	8.2	0	21	260,474	14,391
Horse	equCab2	2007	2,485	76	39	12	27	2	4	1	86	0.5	8.9	4.0	8	21	55,283	46,750
Hedgehog	eriEur2†	2012	2,716	63	28	3	85	2	1	0	86	1.4	8.5	5.3	0	21	219,764	3,265
Cat	felCat8	2014	2,641	34	28	7	22	2	0	0	88	0.7	7.3	4.3	4	21	100,040	18,072
Ferret	musFur1†	2011	2,411	52	26	7	53	2	1	0	89	1.0	6.5	3.8	0	20	109,700	9,335
Microbat	myoLuc2†	2010	2,035	162	36	13	93	7	5	0	83	3.9	9.1	7.4	4	20	61,131	4,293
Sheep	oviAri31	2012	2,619	67	77	14	54	0	0	1	81	1.2	11	7.2	3	21	125,067	100,080
Megabat	pteVam2†	2014	2,198	65	29	17	110	0	0	1	87	0.7	7.7	4.8	4	20	189,339	5,954
Shrew	sorAra2†	2012	2,423	117	33	10	66	6	0	0	85	1.2	7.6	6.9	0	21	188,953	22,794
Pig	susScr102	2011	2,809	210	81	28	213	25	12	1	69	2.4	12	17	7	20	238,439	576
Dolphin	turTru2†	2012	2,552	24	41	23	469	1	3	8	72	1.5	14	13	4	21	313,713	116
Alpaca	vicPac2†	2013	2,172	48	33	19	56	4	0	0	87	1.2	7.9	4.1	0	21	174,225	7,264
Aves																		
Zebra finch	taeGut2	2013	1,232	816	125	15	81	13	12	6	77	2.0	8.7	13	4	20	87,710	8,237
Mallard duck	anaPla1†	2013	1,105	979	117	19	119	11	14	2	72	0.7	10	16	0	20	125,115	1,234
Chicken	galGal4	2011	1,047	686	79	8	50	10	7	0	85	0.9	5.5	8.8	4	20	11,109	12,877
Turkey	melGal5	2014	1,128	637	96	20	376	7	5	2	74	0.5	10	14	0	20	64,955	3,801

Notes: rRNA is the completeness of one 45S ribosomal DNA cluster consisting of pRNA, 28S, 5.8S, and 18S rRNAs in exactly this 5′ to 3′ order. tRNA is the occurrence of 21 amino acids (aa). Gaps are 10 or more nucleotides long. Contiguity is the scaffold N50. PCE, UC, tRNA, and Gaps are absolute counts [#], BUSCO is in percentage [%], rRNA is presented as a score, and genome size and N50 are sequence lengths. The assembly version is the UCSC Genome Browser assembly ID. †Assembly level is scaffold, otherwise chromosome.
Abbreviations: PCE, Protein-coding exons could be D, deleted; P, partially deleted; S, split; or in W, wrong order. UC, Ultraconserved elements could be D, deleted; P, partially deleted; or S, split. BUSCO, Universal single-copy orthologs could be C, complete; CD, complete duplicated; F, fragmented; or M, missing.

these highly conserved genomic features cover both coding and noncoding intergenic regions which is in contrast to what is done in Benchmarking Universal Single-Copy Orthologs (BUSCOs; see below).

Benchmarking Universal Single-Copy Orthologs. Sets of BUSCOs are orthologous groups of single-copy genes described by Simão et al.[20] Any BUSCO in vertebrates can be expected to be found as a single-copy ortholog in any genome from the phylogenetic clade of vertebrates. In short, for each BUSCO group, an amino acid consensus sequence is generated from its respective hidden Markov model profile, and a block profile is built to guide automated gene predictions with `AUGUSTUS`.[25] During genome assessment, regions in a genome that are likely to encode BUSCO-matching genes are identified by tBLASTn searches, then genes are predicted in these candidate regions using the corresponding BUSCO group's block profile and default gene finding parameters. Successful AUGUSTUS gene prediction for each BUSCO group produces an initial BUSCO gene set whose protein sequences are then evaluated using the BUSCO-specific cutoffs to determine true orthology and completeness. Finally, significant matching protein sequences are tested to be likely orthologous or just homologous by applying the BUSCO group's hidden Markov model profile. We were running the BUSCO version 1.1b1 in genome mode with the lineage specific profile libraries of vertebrata and used the precomputed metaparameters of human for placental mammals and chicken for birds (parameter -species human/chicken).

45S ribosomal DNA cluster. Ribosomal RNAs (rRNAs) are the primary structural components of the ribosome. The rRNA species 28S and 5.8S from the large ribosomal subunit and 18S rRNAs from the small subunit are encoded by the 45S ribosomal DNA (rDNA) cluster. Transcription by RNA polymerase I yields a primary transcript (45S pre-rRNA), which is processed into the mature 28S, 18S, and 5.8S rRNAs found in cytoplasmic ribosomes. The rRNAs reciding in a single 45S transcription unit are separated by spacers and are always arranged in the same 5′ to 3′ order: 18S, 5.8S, and 28S. rDNA silencing is mediated through methylation of the rDNA promoter via DNMT3B and pRNA, a noncoding RNA which has been shown to originate from a spacer promoter located upstream of the pre-rRNA transcription start site.[26] We predicted the 28S and 18S rRNAs with `RNAmmer`[27] and searched the 5.8S rRNA Rfam family RF00002 and the pRNA Rfam family RF01518 with `Infernal`.[28] We defined the rRNA score to describe the completeness of the 45S rDNA cluster as $2 \times R - S$, where R is the count of pRNA, 18S, 18S, or 5.8S rRNA in the correct order on the same chromosome or scaffold ($2 < R < 4$), and S is 1 if not all items are located on the same strand and 0 otherwise. The cluster with the highest rRNA score was reported.

Additional features. We counted the presence of tRNAs coding for each of the 20 standard amino acids and selenocysteine and required at least one tRNA coding for each. tRNAs are predicted with tRNAscan-SE.[29] Assembly sequence quality was measured by counting gaps of 10 or more nucleotides in the genome assemblies. The assembly contiguity was described by the scaffold N50 metrics as documented in the NCBI Assembly database (http://www.ncbi.nlm.nih.gov/assembly). Scaffold N50 is a scaffold size such that scaffolds of this length or longer include half of the bases of the assembly.

Genome assembly quality ranking. A quality score for genome assemblies has been previously suggested by combining normalized feature scores.[30] Besides using some of their presented features, we decided to rank the genome qualities without weighting each of the features used to analyze the genomes. The impact of each feature for describing the assembly quality is unknown, and a perfect vertebrate genome assembly to train the weighting parameters does not exist (even the human assembly is still incomplete). Hence, model training would necessarily result in a biased score toward the defined standard. Instead, we decided to measure the differences between the assembly qualities of studied species by reducing the variances of the applied features. This was done by a principal component analysis (PCA) of the features (princomp function from the built-in R stats package). Each genome assembly was represented by a vector consisting of 15 z-score normalized features (see Table 1): highly conserved protein-coding exons (PCE; 4 features), ultraconserved elements (UC; 3 features), universal single-copy orthologs (BUSCO; 4 features), 45S rDNA cluster (rRNA; 1 feature), tRNAs (1 feature), assembly size normalized gap count (1 feature) and contiguity (scaffold N50; 1 feature). Then, the ranking of assembly qualities was quantitatively measured in comparison to the human genome assembly that has been the most intensively investigated of all vertebrate genomes. We calculated the Euclidian distances of the first three principal components (PCA score) between each species and human and ranked the genome qualities accordingly. The number of principal components chosen for the distance measure explained most of the feature variances.

Transcriptome data and processing. For comparison of the read mappability to the genome assembly quality, we downloaded paired-end libraries of polyA-selected RNA and total RNA from the sequence read archive.[31] All libraries were sequenced on an Illumina HiSeq 2000. Only species with at least three libraries were considered. The applied libraries are listed in the Supplementary File 1. For 11 species, we found polyA-selected RNA libraries, and for 4 species, we found total RNA libraries. For human, mouse, and sheep, we studied both polyA-selected and total RNA libraries. We processed the raw reads by removing low-quality reads and adapter sequences with `cutadapt` (version 1.8.3; parameter -m 30 -q 20).[32] Cleaned reads were aligned to their reference genome, which was built without annotations using `STAR` (version 2.4.0 j).[33] After aligning, we removed rRNAs from the mapped transcriptomic data based on the rRNA predictions from `RNAmmer` (8S, 18S, and 28S) and `Infernal` (Rfam families RF00001:5S and RF00002:5.8S). We counted uniquely and

multimapped reads as mapped reads. For each organism, we documented the mean and standard deviation of mapped reads in the applied libraries.

Results

In this study, we analyzed the genome quality of the latest assemblies (September 2015) of 20 domesticated and phylogentically related animals from the classes Laurasiatheria (placental mammals) and Aves (birds), including several farm animals. As a gold standard assembly, the human genome (hg19) has been included. The phylogenetic relationship between the species is shown in Figure 1A.

Genome assembly quality features. The analyzed genome assembly quality features of all 21 species are summarized in Table 1. At a first glance, we see that most of the applied features have their best values for the human and mouse genome assemblies, which is in agreement with the extensive efforts undertaken to study these organisms. The efforts to complete the genomes are especially reflected in the gap content, which is much lower for human and mouse than in the other species. Another strong signal is that, in general, features for UCs, conserved PCEs, and universal single-copy orthologs (BUSCOs) are of lowest quality for the bird genome assemblies, which may be partly explained by the evolutionary distance to mammals. However, chicken has arguably an assembly quality better than that of many mammals, and this is likely due to the usage of the Sanger sequencing technology. Nine genome assemblies are still at the scaffold level; however, we do not see a clear quality difference to the assemblies at the chromosome level.

Strikingly, all genome assemblies lack a significant amount of the 3,023 BUSCOs. Human, mouse, dog, and ferret have the largest number of complete BUSCOs (89%–91%). The least complete genomes in terms of BUSCOs are those of pig, dolphin, mallard duck, turkey, zebra finch, and domestic goat (69%–79%). The sequence conservation and synteny of 4,856 PCEs generally agree with the BUSCO assessment. However, an exception is cow, which performs very well in terms of PCEs but less well in BUSCOs. The genome-wide alignment based comparisons to human (PCEs) are likely to perform better than BUSCOs because synteny with human had been used in the built of the cow assembly.[11] The 473 UCs are well covered by all genomes. An exception of this trend is the pig assembly with 8% incomplete UCs, which is more than that for the genomes of birds. It has been suggested that during the initial pig genome project, only about 90% of the pig genome was accessible in BAC clones,[12] which could explain the incomplete set of intergenic elements in our study.

Extreme cases of assembly contiguity are the sheep, pig, and dolphin assemblies. The scaffold N50 of sheep is very high (100,080 Kbp), whereas the contig N50 is much lower (40,376 Kbp), suggesting issues in the scaffolding. In contrast, the scaffold N50 of pig and dolphin is very low, which is in agreement with the low quality of these two assemblies described by the other features. The 45S rDNA cluster is a highly repetitive genomic region that makes it hard to assemble in the correct order without very long reads. Besides human, only the genome assemblies of dog and horse have a complete 45S rDNA cluster. All the other genomes completely lack the cluster or contain only a part of it. All genomes have a complete set of standard tRNA codons and only part of them, including human, miss a codon for selenocysteine.

The pig genome is clearly the least complete assembly of all placental mammals. It misses a substantial number of UCs and a large fraction of PCEs. On the other hand, it is one of the few genomes with a (almost) complete 45S rDNA cluster. 28S, 5.8S, and 18S are located in a tandem on chromosome 6, but the rDNA silencing mediator pRNA is positioned almost 100 kb upstream on the opposite (positive) strand. This suggests that the pig genome has been exhaustively assembled on highly incomplete genomic sequencing. The dolphin genome assembly represents another extreme with almost 10% of the PCEs in a rearranged order in comparison to human, which may be partly explained by the large amount of scaffolds (240,901; see also low scaffold N50). The genome assembly of panda illustrates that the features of genome sequence quality and gene content do not agree in all cases. Whereas the low scaffold N50 and the large number of gaps suggest a low quality, the gene content-based metrices are among the best of all examined assemblies (except of the rRNA feature). Hence, in the following, we propose a ranking of genome assembly quality combining all features.

Quantitative ranking of genome assembly qualities. For the 15 quality features, the first three principal components (PCs) account for 75% of their variance. Figure 2 illustrates important relationships between the assessed features and the three PCs. Almost 50% of the information in the features is reduced into the first PC (PC1), which is primarily composed of features describing misassembled protein-coding genes or UCs (fragmented, split, wrong order). Features describing missing genomic information are primarily represented by PC2 (deleted, partially deleted), and this accounts for 17.2% of the variance. PC3 describes another 10.8% of the feature variance that originates mostly from the complete duplicated BUSCO and the 45S rDNA cluster. The assembly contiguity (N50) cannot be grouped with the other features and, hence, contributes equally to all three PCs.

Based on these three PCs, we quantitatively rank the species by their Euclidian distance to human in the three-dimensional space, see PCA score in Figure 3. The PCA score can be interpreted as an assembly quality score. The genome assemblies of dog and mouse are of highest quality followed by cow, horse, cat, ferret, and microbat. The genomes of sheep, megabat, hedgehog, shrew, panda, alpaca, and chicken are all of medium quality, and the genomes of dolphin, mallard duck, pig, turkey, zebra finch, and domestic goat have the lowest PCA scores. Figure 3A shows a weak positive correlation between the PCA score and the divergence time between human and the compared species (Pearson's correlation

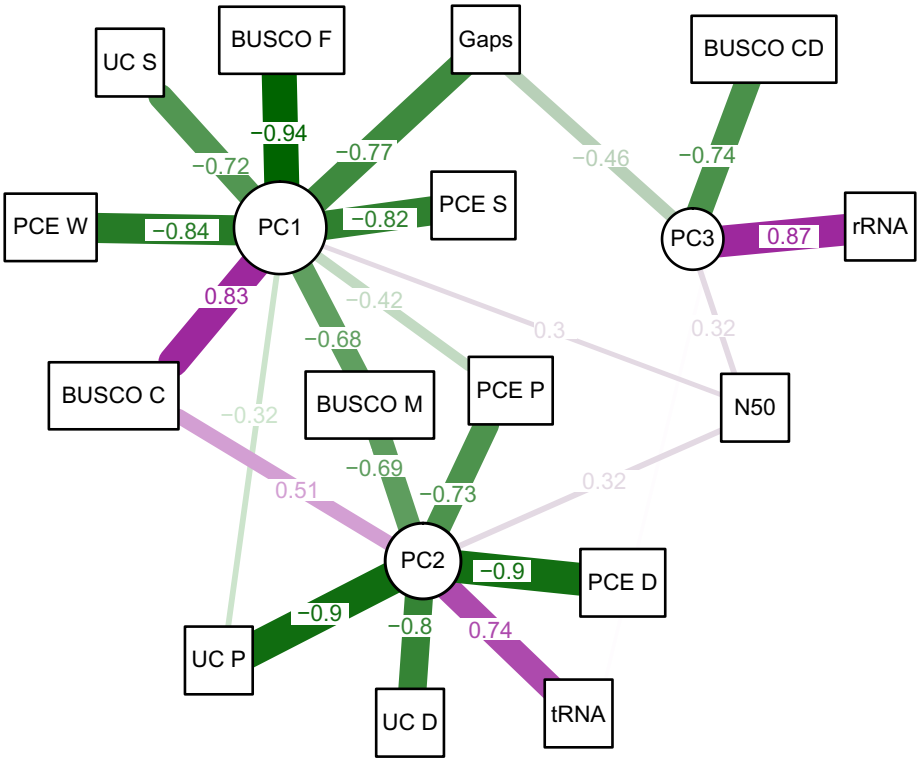

Figure 2. Relationship between principal components and quality features.
Notes: The first three principal components (PCs) account for 75% of the feature variance (PC1: 47.1%, PC2: 17.2%, and PC3: 10.8%). Rectangular nodes describe the 15 quality features. The edge weight describes how much variance of a feature is explained by the principal component. Green edges connect features negatively related to genome quality, and purple edges connect features positively related to genome quality. Relations (edges) are shown if greater than 0.3 or smaller than −0.3. See Table 1 for abbreviations of the quality features. Gaps are normalized by assembly size. The figure was made using the R qgraph package.

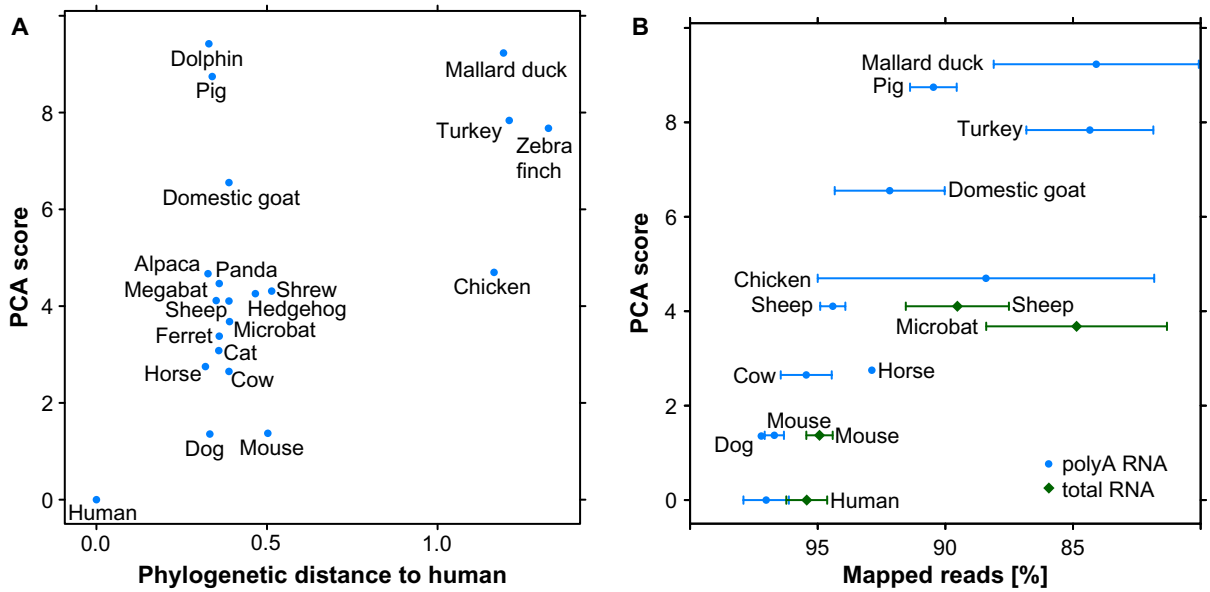

Figure 3. Correlation of genome assembly quality to (**A**) phylogenetic distance to human and to (**B**) frequency of mapped reads from RNAseq experiments.
Notes: The genome assembly quality (PCA score) is measured as the Euclidian distance of principal component 1 (PC1), PC2, and PC3 between human and 20 other species. The human genome serves as reference and has a Euclidian distance of zero. RNAseq experiments are divided into polyA-selected RNA (blue circle) and total RNA (green diamond), and the mean and standard deviation of mapped reads are shown. After removing human (reference) and bird genomes (large phylogenetic distance), the Pearson's correlation coefficient between assembly quality and phylogenetic distance is 0.26, between assembly quality and polyA-selected RNA mapped reads is 0.91, and between polyA-selected RNA mapped reads and phylogenetic distance is 0.43. Only the correlation between assembly quality and polyA-selected RNA mapped reads is significant ($P < 0.005$).

coefficient $\rho = 0.54$). However, after removing human (gold standard) and the bird genomes (which have large phylogenetic distance to the placental mammals), the correlation disappears ($\rho = 0.26$). Given that the mouse is more evolutionarily distant to human than any of the other mammals considered, and its genome assembly is of high quality, we note that the presented quality score is not biased by phylogeny.

Mappability of sequencing data. Herein, we address how strongly the genome assembly quality impacts the mappability of RNAseq data to the reference genome (see Supplementary File 1). That is, we are interested in knowing how much information we loose in NGS studies merely due to a suboptimal genome assembly quality. In Figure 3B, we show a positive correlation between assembly quality (measured by the PCA score) and the percentage of mapped reads from polyA-selected RNA libraries (Pearson's correlation coefficient $\rho = 0.86$; *t*-test $P < 0.001$), which indicates the importance of high-quality genome assemblies for maximal gain from NGS data. Without the human and the bird genomes, the Pearson's correlation coefficient between mapped reads (polyA RNA) and assembly quality is even higher ($\rho = 0.91$; *t*-test $P < 0.005$), whereas the correlation between mapped reads (polyA RNA) and the evolutionary distance to human is not significant ($\rho = 0.43$). The trend for total RNA is similar, but a correlation analysis was not possible due to the small number of total RNA libraries in this study. However, the data suggest that the number of mapped reads depends even more on assembly quality for total RNA libraries in comparison to polyA-selected RNA libraries.

Cases of missing genotypes in pig. The varying assembly quality of domesticated animals may have a large impact on pathway reconstruction due to the missing genes. Below, we describe two examples in pig, which is an important production animal as well as a useful model organism.

The first example is the DGAT2 gene, which codes for an enzyme that catalyzes triglyceride synthesis[34] in eukaryotes. The gene does not have an annotated ortholog in pig. In our study, the `lastz` pairwise alignment to human aligns four exons of the human DGAT2 homolog to the DGAT2-like 6 gene in pig, but in the 5′ end of the gene, three of the exons are aligned to three different chromosomes. However, the gene has been isolated in pig by cDNA cloning procedures.[35] The polymorphisms of the gene play a role in backfat tissue quality, which is an important trait for the meat product industry.[34–36] DGAT2 is also of interest in the study of obesity in humans, and it has been shown to be upregulated in obese pigs, which are used as model organisms for obesity.[37] Hence, missing this gene in systems biology analyses could have unfortunate consequences.

The second example is the cholesteryl ester transfer protein (CETP), a protein playing a central role in atherosclerosis, the chronic inflammatory condition causing most cardiovascular diseases[4] and therefore a leading cause of death worldwide.[38] High levels of low-density lipoproteins (LDL) and low levels of high-density lipoproteins (HDL) play a main role in atherosclerosis,[39] and the cholestryl ester transfer protein is specifically the one responsible for controlling HDL-to-LDL ratios.[40] Pig is a suitable model organism in the study of atherosclerosis, due to the spontaneous occurrence of the disease, size, and its human-like cardiovascular anatomy.[3,4] However, the CETP gene is not annotated in pig, and it is not clear whether this is a genome assembly issue, or whether the gene is really not present in the animal. The gene is known to be naturally lacking in mouse,[4] despite being present in other mammals such as rabbit or dolphin. A genome analysis study performing de novo genome assembly in mini pig concluded that CETP was among the genes lost in the lineage,[41] while the authors of a previous study have supposedly cloned the gene in pig.[42] The low levels of the protein, detected in pig by antibody designed against the human CETP,[43] could be explained by the presence of an inhibitor of the protein, a hypothesis supported by a study where a human CETP inhibitor was isolated from pig plasma.[44] Due to the low quality of the pig genome, we cannot draw final conclusions about the existence or about the genotypes of these obviously important genes for meat production and disease modeling. In addition, genetic analyses of the respective epigenetic and regulatory marks of these genes are, therefore, not possible to be performed in pig.

Discussion

A key result of our study is the urgent need to reinforce the efforts for improving genome assembly quality in domesticated animals. Using a variety of different quality features, we show that many of the investigated genome assemblies are far from perfect, characterized by missing or fragmented vertebrate-wide conserved genomic loci and low scaffold contiguity. The low cost of short-read NGS-based sequencing has boosted the sequencing of domesticated animal genomes and transcriptomes, among other species. The consequences of poor genome assembly quality become most obvious in the mappability of NGS reads. While we lack enough total RNA-sequencing data for domesticated animals, we observe a clear trend of lower mappability of polyA-selected RNAseq in lower quality assemblies. Genome and transcriptome annotations are largely affected by missing or fragmented genomic content that may lead to wrong conclusions about the genes or transcripts present in the organism. Also comparative genetics relies on correctly sequenced, aligned, and annotated genomes, and we have shown the possible issues in two pig examples.

The presented quality measure focuses on contiguity and completeness of genome assemblies. The human genome is the most studied, and hence, we used its assembly as a gold standard for characterizing the completeness of the other assemblies. Ideally, the genome quality should be exclusively based on independent features without a gold standard genome because the genomic difference between human and the analyzed species may introduce a phylogenetic bias. To increase the feature space in this study, we decided to include human-based completeness measures, and the high-quality score we obtained for the mouse

genome illustrates the usability of the presented approach to quantify assembly quality. In addition, we used the human assembly to rank the quality of the species assemblies, which, however, has no impact on the measured quality features.

We showed that traditional Sanger sequencing, characterized by longer read length and better quality, led to better assembly quality than short-read NGS-based sequencing. Dog, horse, or cow, all three Sanger based, are top scoring according to our ranking, while sheep and goat, based on short-read NGS, are of worse quality. The BACs used in the pig genome assembly, another assembly of low quality, were sequenced using Sanger technology, whereas the gaps between BACs were closed using short-read NGSs. Sheep has very high NGS coverage and its genome has been iteratively refined, which is reflected by higher scoring than both goat and pig. However, high coverage of short-read NGS is rare enough to achieve high-quality assemblies. This is primarily due to the repetitive content of the genomes, including repetitive DNA near centromeres and telomeres, large paralogous gene families, and retrotransposons such as LINEs and SINEs.

An important step toward increased quality and usability of the genomes is the incorporation of new data based on long-read sequencing and mapping technologies. Most recently, long-range sequencing has been dramatically improved by Pacific Biosciences (PacBio) Single Molecule Real Time and Oxford Nanopore, and mapping by the Dovetail Genomics Chicago protocol and the 10X Genomics Chromium instrument. For example, the PacBio RS II technology updated in 2014 is advertised as producing raw reads with mean lengths of 15 kb at the cost of error rates as high as 15% and about 100-fold higher expenses than short-read NGS.[45] However, per-nucleotide accuracy of 99.99% can be achieved through algorithmic techniques and sufficient coverage.[7] Not surprisingly, several genome assemblies are currently complemented with PacBio sequencing, such as chicken (galGal5) and sheep (oviAri4). Mapping technologies improve scaffold contiguity and synteny by determining the long-range information on the arrangement of DNA without sequencing every base. For example, the Dovetail Genomics Chicago protocol,[46] introduced in spring 2015, studies the 3D contacts of *in vitro* reconstituted chromatin through an optimized Hi-C approach. It can achieve DNA spanning up to ~150 kb length and has been successfully applied to improve the existing assembly of the American alligator.

Conclusion

The genomes and transcriptomes of domestic animals deserve optimal exploration for making improvements in productivity without compromising animal welfare, as well as for studying human genetics and diseases. The analyses of a comprehensive list of genome assembly features of domesticated animals and related species illustrate the large discrepancy between their assembly quality and NGS efforts. Especially the farm animals pig, chicken, sheep, and cow, which are of high economical and ecological importance, lack a significant number of core eukaryotic and universal genes in their current genome assemblies. Our study presents a novel way of ranking the assembly qualities in comparison to a gold standard. The data and pipeline presented in this study can be applied to judge the assembly quality and the number of unmapped reads in a NGS study. We show that the exploitation rate of RNAseq data is correlated with the genome assembly quality. We conclude that more efforts are needed to improve the genome assemblies of domestic animals. Especially due to the affordable access to the aforementioned new technologies, we expect a significant improvement in the quality of domesticated animal genomes in the near future.

Acknowledgment

We thank Caroline Junker Mentzel for her comments to this article. We would also like to thank Lars Juhl Jensen for his input in particular about the coverage of highly conserved genes.

Author Contributions

Conceived and designed the experiments: SES. Analyzed the data: SES, CA. Wrote the first draft of the manuscript: SES. Contributed to the writing of the article: SES, JG, OP, CA. Agreed with manuscript results and conclusions: SES, CA, OP, JG. Jointly developed the structure and arguments for the paper: SES, JG. Made critical revisions and approved the final version: SES, CA, OP, JG. All the authors reviewed and approved the final article.

Supplementary Material

Supplementary File 1. RNAseq libraries used to assess genome assembly quality and read mappability to the reference genome.

REFERENCES

1. Spurlock ME, Gabler NK. The development of porcine models of obesity and the metabolic syndrome. *J Nutr*. 2008;138:397–402.
2. Mentzel CM, Anthon C, Jacobsen MJ, et al. Gender and obesity specific microRNA expression in adipose tissue from lean and obese pigs. *PLoS One*. 2015;10:e0131650.
3. Vilahur G, Padro T, Badimon L. Atherosclerosis and thrombosis: insights from large animal models. *J Biomed Biotechnol*. 2011;2011:907575.
4. Getz GS, Reardon CA. Animal models of atherosclerosis. *Arterioscler Thromb Vasc Biol*. 2012;32:1104–15.
5. Swindle MM, Makin A, Herron AJ, Clubb FJ Jr, Frazier KS. Swine as models in biomedical research and toxicology testing. *Vet Pathol*. 2012;49:344–56.
6. Gutierrez K, Dicks N, Glanzner WG, Agellon LB, Bordignon V. Efficacy of the porcine species in biomedical research. *Front Genet*. 2015;6:293.
7. Lee H, Gurtowski J, Yoo S, et al. Third-generation sequencing and the future of genomics. *bioRxiv*. 2016.
8. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. *Genome Res*. 2010;20:1165–73.
9. Lindblad-Toh K, Wade CM, Mikkelsen TS, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*. 2005;438:803–19.
10. Wade CM, Giulotto E, Sigurdsson S, et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*. 2009;326:865–7.
11. Zimin AV, Delcher AL, Florea L, et al. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol*. 2009;10:R42.
12. Groenen MA, Archibald AL, Uenishi H, et al. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*. 2012;491:393–8.
13. Jiang Y, Xie M, Chen W, et al. The sheep genome illuminates biology of the rumen and lipid metabolism. *Science*. 2014;344:1168–73.
14. Dong Y, Xie M, Jiang Y, et al. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nat Biotechnol*. 2013;31:135–41.

15. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*. 2004;432:695–716.
16. Dalloul RA, Long JA, Zimin AV, et al. Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol*. 2010;8.
17. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29:1072–5.
18. Clark SC, Egan R, Frazier PI, Wang Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*. 2013;29:435–43.
19. Parra G, Bradnam K, Ning Z, Keane T, Korf I. Assessing the gene space in draft genomes. *Nucleic Acids Res*. 2009;37:289–97.
20. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31:3210–2.
21. Harris RS. Improved Pairwise Alignment of Genomic DNA [Ph.D. thesis]. The Pennsylvania State University; 2007. Available at: http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf
22. Kent J. UCSC Tools; 2015. Available at: http://genome.ucsc.edu/index.html.
23. Ciccarelli FD, Doerks T, von MC, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. *Science*. 2006;311:1283–7.
24. Bejerano G, Pheasant M, Makunin I, et al. Ultraconserved elements in the human genome. *Science*. 2004;304:1321–5.
25. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*. 2011;27:757–63.
26. Schmitz KM, Mayer C, Postepska A, Grummt I. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev*. 2010;24:2264–9.
27. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*. 2007;35:3100–8.
28. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. 2013;29:2933–5.
29. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 1997;25:955–64.
30. Land ML, Hyatt D, Jun SR, et al. Quality scores for 32,000 genomes. *Stand Genomic Sci*. 2014;9:20.
31. Kodama Y, Shumway M, Leinonen R; International Nucleotide Sequence Database Collaboration. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res*. 2012;40:D54–6.
32. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011;17:10–2.
33. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21.
34. Yue H, Biao L, JiYing W, Ying W, ShuDong W, YunLiang J. Analysis on the developmental expression of Acyl-coA: diacylglycerol acyltransferase genes (DGAT1 and DGAT2) in porcine backfat tissues. *Nong Ye Sheng Wu Ji Shu Xue Bao*. 2010;18:905–10.
35. Yin Q, Yang H, Han X, Fan B, Liu B. Isolation, mapping, SNP detection and association with backfat traits of the porcine CTNNBL1 and DGAT2 genes. *Mol Biol Rep*. 2012;39:4485–90.
36. Renaville B, Bacciu N, Lanzoni M, Corazzin M, Piasentier E. Polymorphism of fat metabolism genes as candidate markers for meat quality and production traits in heavy pigs. *Meat Sci*. 2015;110:220–3.
37. Jacobsen MJ, Mentzel CM, Olesen AS, et al. Altered methylation profile of lymphocytes is concordant with perturbation of lipids metabolism and inflammatory response in obesity. *J Diabetes Res*. 2016;2016:8539057.
38. Weber C, Noels H. Atherosclerosis: current pathogenesis and therapeutic options. *Nat Med*. 2011;17:1410–22.
39. Badimon L, Vilahur G. LDL-cholesterol versus HDL-cholesterol in the atherosclerotic plaque: inflammatory resolution versus thrombotic chaos. *Ann NY Acad Sci*. 2012;1254:18–32.
40. Barkowski RS, Frishman WH. HDL metabolism and CETP inhibition. *Cardiol Rev*. 2008;16:154–62.
41. Fang X, Mou Y, Huang Z, et al. The sequence and analysis of a Chinese pig genome. *Gigascience*. 2012;1:16.
42. Shi XW, Zhang YD, Rothschild MF, Tuggle CK. Rapid communication: genetic linkage and physical mapping of the porcine cholesteryl ester transfer protein (CETP) gene. *J Anim Sci*. 2002;80:1390–1.
43. Speijer H, Groener JE, van RE, van TA. Different locations of cholesteryl ester transfer protein and phospholipid transfer protein activities in plasma. *Atherosclerosis*. 1991;90:159–68.
44. Cho KH, Lee JY, Choi MS, Bok SH, Park YB. Interaction of CETP inhibitory peptide and lipoprotein substrates in cholesteryl ester transfer assay: relationship between association properties and inhibitory activities. *Lipids*. 2002;37:641–6.
45. Koren S, Schatz MC, Walenz BP, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol*. 2012;30:693–700.
46. Putnam NH, O'Connell BL, Stites JC, et al. Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res*. 2016;26:342–50.